Geodatenfusion im Kontext von GDI und Semantic Web

STEFAN WIEMANN[1]

Zusammenfassung: Die Verknüpfung von Geodateninfrastrukturen mit Semantic Web Technologien kann einen bedeutenden Beitrag zur Etablierung einer ubiquitär verfügbaren und informationsreichen Geodatenbasis im Web darstellen. Methoden zur Geodatenfusion spielen dabei eine große Rolle, da sie beide Entwicklungen überbrücken und einen echten Informationsmehrwert für den Anwender schaffen können. Diese Arbeit untersucht Potentiale und Möglichkeiten einer entsprechenden Umsetzung im Projekt COBWEB. Unter anderem sollen darin durch Crowdsourcing gesammelte Umweltbeobachtungen mit administrativen Grenzen fusioniert und als Linked Data zur Verfügung gestellt werden. Auf dieser Grundlage können im Anschluss nutzerspezifische Abfragen über die verlinkten Geodatenquellen ausgeführt werden.

1 Einleitung

Die Entwicklung des Semantic Web und der zunehmende Wandel von daten- hin zu dienstebasierten Strukturen im Internet hat die Bedeutung und öffentliche Wahrnehmung von Geodaten bereits grundlegend verändert. Aus der online verfügbaren, kontinuierlich wachsenden Menge an Geodaten nutzbare Informationen zu generieren, stellt jedoch weiterhin eine große Herausforderung dar. Dazu zählt auch die Fusion verfügbarer Daten über Webdienste, die einen Mehrwert durch die dynamische Verlinkung und Kombination entsprechender Datensätze oder der darin enthaltenen Objekte schaffen kann.

Die Umsetzung einer dienstebasierten Fusion von Geodaten erfolgt in der Regel über eine Sequenz mehrerer Teilprozesse (z.B. Datenharmonisierung, Datenvergleich, Zusammenführung von Daten, Fehlerkorrektur), die je nach Anwendungsfall unterschiedlich kombiniert werden können. Um dies zu ermöglichen, werden offene Standards zur Datenrecherche, -bereitstellung und -prozessierung sowie Verlinkung von Geodaten benötigt. Erstere können im Wesentlichen durch existierende Standards des Open Geospatial Consortium (OGC) abgedeckt werden. Für die Verlinkung von Daten eignet sich dahingegen die Anwendung des Linked Data Paradigma, einem der Grundbausteine des Semantic Web (BIZER et al. 2006). Ein Teilergebnis des Fusionsprozesses sowie Grundlage einer Zusammenführung von Geodaten ist entsprechend eine Vielzahl an Verknüpfungen zwischen existierenden Datenbeständen, die neben verlinkten Objekten auch weiterführende Informationen enthalten können. Als Anwendungsbeispiel in dieser Arbeit dient die Fusion von freiwillig gesammelten Umweltbeobachtungen mit administrativ erhobenen Datensätzen im Kontext des Projektes COBWEB[1] (Citizen Observatory Web). Darin stehen die Kombination und Weiterverarbeitung erfasster Daten mit bereits existierenden Bestandsdatensätzen im Vordergrund. Eine wichtige Rolle spielt zudem die Verwaltung von Qualitäts- und Konfidenzmaßen im Fusionsprozess.

1) Stefan Wiemann, Professur für Geoinformationssysteme, Technische Universität Dresden, Helmholtzstr. 10, 01069 Dresden; E-Mail: stefan.wiemann@tu-dresden.de

[1] http://cobwebproject.eu/

Als prototypische Implementierung dient die Umsetzung des zuvor genannten Anwendungsfalles unter Nutzung von OGC Web Processing Services (WPS, OGC 2007) und Linked Data. Im Ergebnis werden verschiedene Sichten auf die Verlinkten Geodaten ermöglicht, welche je nach Anwendungsfall die wesentlichen Informationen aus dem Fusionsprozess abbilden.

2 Das Projekt COBWEB

Das von der Europäischen Union im Rahmen des Framework Programme 7 (FP7) geförderte Projekt COBWEB (Citizen Observatory Web) beschäftigt sich mit der Entwicklung eines Online-Informationssystems zur kollaborativen Sammlung und Analyse von Umweltdaten in UNESCO Biosphärenreservaten. Gesammelte Umweltinformationen sollen Entscheidungsträgern bei der Gestaltung und Umsetzung politischer Vorhaben unterstützen und somit einen Beitrag zur nachhaltigen Entwicklung der Biosphärenreservate leisten.
Landschaften können den Status eines UNESCO Biosphärenreservates erlangen, indem sie einen entsprechend starken Rückhalt in der Gemeinschaft besitzen und nachweislich ausreichend Kapazitäten aufweisen, um die drei Funktionen Schutz (Beitrag zur Erhaltung von Landschaften, Ökosystemen, Arten und genetischer Vielfalt), Entwicklung (Förderung einer wirtschaftlichen und menschlichen Entwicklung, die soziokulturell und ökologisch nachhaltig ist) und logistische Unterstützung (Förderung von Demonstrationsprojekten, Umweltbildung und -ausbildung, Forschung und Umweltbeobachtung im Rahmen lokaler, regionaler, nationaler und weltweiter Themen des Schutzes und der nachhaltigen Entwicklung) zu erfüllen.
Die ausgewählten Testgebiete für das Projekt sind im Wesentlichen das walisische Biosphärenreservat *Dyfi*, das deutsche Biosphärenreservat *Schleswig-Holsteinisches Wattenmeer und Halligen* sowie die griechischen Biosphärenreservate *Olymp* und *Samaria-Schlucht*. Das Ziel ist die Beteiligung der jeweiligen Bevölkerung und Besucher im Biosphärenreservat zu steigern, um Entscheidungsprozesse vor Ort zu verbessern sowie robuste Methoden und Technologien zu entwickeln, die anschließend auch in anderen Anwendungen hilfreich sein können. Als Schwerpunkte des Projektes wurden die Validierung von Fernerkundungsdaten, Beobachtung von Flora und Fauna sowie Hochwasserereignisse ausgewählt.
Das grundlegende Prinzip, auf dem das Projekt beruht, wird vielfach als Crowdsourcing bezeichnet. Im Allgemeinen beschreibt dies die Sammlung und Veröffentlichung (sourcing) von, bislang von ausgebildeten Fachleuten erhobenen, Daten durch freiwillige Helfer oder eine Gemeinschaft (crowd), deren Motivation und Fähigkeiten zunächst unbekannt sind. Im Kontext geographischer Daten hat sich für dieses Prinzip der Begriff *Volunteered Geographic Information* herausgebildet (GOODCHILD 2007).
Die Entwicklungsschwerpunkte von COBWEB liegen auf der Verbesserung von Möglichkeiten zur Datenerfassung über mobile Endgeräte, der Qualitätssicherung erfasster Daten sowie der Operationalisierung standardisierter Geodateninfrastrukturen. COBWEB soll die Aggregation und Fusion gesammelten Daten mit Bestandsdaten unterstützen und diese in geeigneter standardisierter Form verfügbar machen. Innerhalb des GEOSS (Global Earth Observation System of Systems) Frameworks wird das Projekt an der Etablierung gemeinsamer Methoden und Standards in den Bereichen Datenarchivierung, -suche und -zugriff arbeiten. Die während

der Projektlaufzeit gesammelten Daten werden entsprechend ohne Einschränkungen, lediglich unter dem Vorbehalt des Datenschutzes, zur Verfügung gestellt.

3 Datenfusion in Geodateninfrastrukturen

Der Begriff der Datenfusion kann je nach Anwendungsgebiet unterschiedlich interpretiert und definiert werden. Im Folgenden beschreibt er die Zusammenführung von Geodaten oder Geoobjekten aus beliebigen Quellen, aus denen ein informatorischer Mehrwert für die jeweilige Anwendung generiert werden kann. Datenfusion im Kontext von Geodateninfrastrukturen kann als zusätzliche Abstraktionsebene zwischen Datendiensten und Anwenderschicht gesehen werden (Abb. 1).

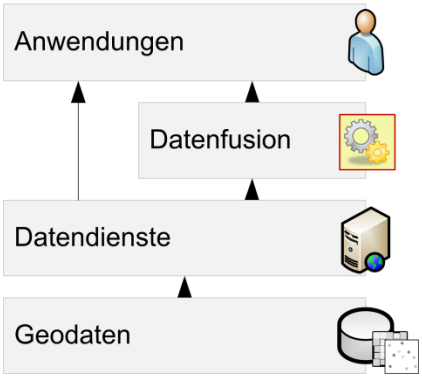

Abb. 1: Datenfusion in Geodateninfrastrukturen

3.1 Grundlagen

Die Mehrheit existierender Anwendungen zur Geodatenfusion sind komplexe, in sich geschlossene Systeme mit konkretem Anwendungsbezug. Obwohl dies für einzelne Anwendungsfälle sehr gut geeignet ist, erfordert die dienstebasierte Geodatenfusion eine Dekomposition in wohl definierte, standardisierte Teilprozesse, die je nach Anwendungsfall unterschiedlich miteinander kombiniert werden können. Dafür wird der Gesamtprozess zunächst in eine Sequenz von sieben Schritten zerlegt (Abb. 2):

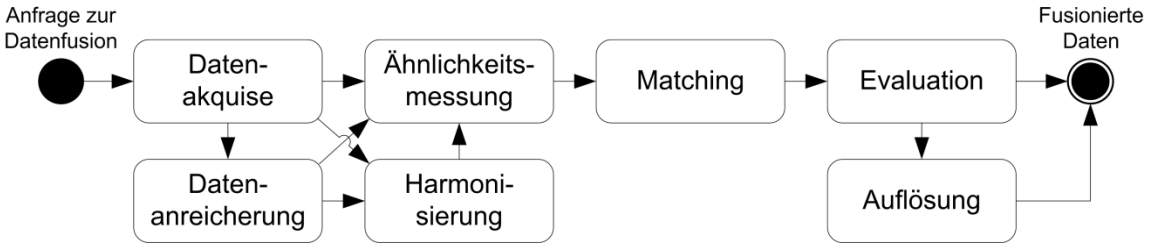

Abb. 2: Sequenz von Teilprozessen einer dienstebasierte Geodatenfusion

1. Datenakquise – regelt die Beschaffung von Eingangsdaten auf Grundlage existierender Standards für die Katalogisierung und Suche, sowie Bereitstellung von Geodaten über das Internet.

2. Anreicherung – adressiert die Charakteristik einzelner Eingangsdaten und beinhaltet Methoden zur Qualitätsprüfung und Fehlerkorrektur. Das Ziel ist die Bereitstellung eines konsistenten und geeigneten Datensatzes für die weitere Prozessierung.
3. Harmonisierung – sorgt für die Angleichung von Eingangsdaten, um einen gewissen Grad an Vergleichbarkeit herzustellen. Beispiele dafür sind die Eliminierung eines räumlichen Versatzes, die Anwendungen von Koordinatentransformation oder Methoden der Modellgeneralisierung.
4. Ähnlichkeitsmessung – bildet die Grundlage für die Datenfusion und ermittelt Vergleichsmaße zwischen Datensätzen und darin enthaltenen Informationen. Diese können sowohl über Geometrien und Attribute von Geoobjekten als auch deren Struktur und beschreibende Schemata berechnet werden.
5. Matching – basiert auf den zuvor berechneten Vergleichsmaßen und ermittelt die Wahrscheinlichkeit von Geoobjekten homolog, also Repräsentationen des gleichen realweltlichen Objektes oder Phänomens, zu sein. Die Bewertung des Matching kann mit zugehörigen Konfidenzmaßen unterlegt werden.
6. Evaluierung – bewertet die Ergebnisse des Matching-Prozesses hinsichtlich auftretender Unterschiede zwischen einander zugeordneten Geoobjekten. Dabei werden alle Änderungen oder Konflikte identifiziert, die im finalen Prozessierungsschritt gelöst werden müssen.
7. Auflösung – beinhaltet je nach Anwendungsfall verschiedene Strategien zur Konfliktlösung, die angewendet werden können, um Ergebnisse entsprechend der Nutzeranforderungen bereitzustellen. Dies kann eine konsistente Zusammenführung von Daten, die Erstellung von Differenzbildern oder den Transfer von Geometrien oder Attributen beinhalten.

Die genannten Teilprozesse werden in einer Umsetzung in der Regel nicht explizit trennbar sein, da es je nach Umsetzungsstrategie sinnvoll sein kann, mehrere von ihnen zusammenzufassen. Zudem können einzelne Teilprozesse optional, iterierbar oder in einer anderen Reihenfolge aneinander gekoppelt sein.

Um einen interoperablen Zugriff und die Verkettung der genannten Funktionalitäten zu ermöglichen, müssen alle Komponenten dem Prinzip einer Serviceorientierten Architektur, im Wesentlichen der Standardisierung, losen Kopplung, Wiederverwendbarkeit, Statuslosigkeit und Komponierbarkeit, folgen (ERL 2008). Im Kontext der Standardisierung von Geodateninfrastrukturen bietet sich hierbei die Nutzung des OGC Web Processing Service Standards an.

3.2 Nutzung von Linked Data Technologien zur Geodatenfusion

Die bisherige Entwicklung von GDI stützt sich im Wesentlichen auf die Bereitstellung, Visualisierung und zunehmend auch Prozessierung von Geodaten. Auf der anderen Seite führt die Entwicklung des Semantic Web zu einer ubiquitär verfügbaren, untereinander verknüpften Datenbasis im Internet. Obwohl beide Entwicklungen unterschiedliche Ziele und Umsetzungsstrategien verfolgen, birgt deren Integration ein großes Potential, insbesondere im Bereich der Geodatenfusion. Die Kombination von dienstebasierten Prozessierungskapazitäten mit Linked Data Technologien zur Verknüpfung von Geodaten kann einen Beitrag zur

Etablierung eines Geospatial Semantic Web leisten, welches derzeitigen GDI-Entwicklungen in Form und Funktionalität in weiten Teilen überlegen wäre (EGENHOFER 2002).
Die Verknüpfung von GDI und Semantic Web im Bereich der Geodatenfusion stellt eine Reihe von Anforderungen, um die gewünschten Vorteil daraus zu ziehen. Dies betrifft im Wesentlichen ein standardisiertes Vokabular für die Beschreibung von Geodaten und Relationen, die Standardisierung räumlicher Abfragen aus verlinkten Ressourcen im Semantic Web sowie Möglichkeiten zur Mediation der unterschiedlichen Datenstrukturen zwischen den Systemen.
Ein standardisiertes Vokabular ist wichtig, um Geodaten und Relationen zwischen Geoobjekten formal und allgemeingültig beschreiben zu können. Im Bereich Geodaten kann hierbei auf bereits existierenden Vokabularen, beispielsweise dem Geospatial Vocabulary[2] des World Wide Web Consortium, der NeoGeo Vocabulary Specification[3] oder dem GeoSPARQL[4] Vokabular des OGC, aufgebaut werden. Je nach Anwendungsgebiet kann ein vorhandenes Vokabular verwendet, gegebenenfalls ergänzt oder weiter spezifiziert werden. Das Gleiche gilt für raumbezogene Erweiterungen der Abfragesprache SPARQL (SPARQL Protocol And RDF Query Language), für die insbesondere der GeoSPARQL Standard des OGC (OGC 2012) von Bedeutung sein wird.
Um die unterschiedlichen Datenstrukturen in traditionellen GDI und Semantic Web zu überbrücken, gibt es im Wesentlichen drei Möglichkeiten: (1) die Transformation von Geodatenbeständen nach Linked Data, (2) die Mediation von Geodatendiensten über entsprechende Transformationsdienste und (3) die direkte Verlinkung von Geodatendiensten innerhalb des Semantic Web. Jede der genannten Möglichkeiten bringt jeweils Vor-, aber auch Nachteile mit sich, die je nach Anwendungsfall gegeneinander abgewogen werden müssen. Wichtige Faktoren bei dieser Entscheidung sind die notwendige Performanz, Sicherheitsbeschränkungen, die Flexibilität existierender Strukturen oder der Synchronisationsbedarf einer Anwendung.

3.3 Umsetzung im Projekt COBWEB

Um die Kombination der dienstebasierten Geodatenfusion mit Linked Data Technologien zu demonstrieren, wurde im Rahmen des COBWEB Projektes ein Prototyp zur Verlinkung von Umweltbeobachtungen mit administrativen Einheiten implementiert. Der Anwendungsfall ist in Abb. 3 skizziert und beschreibt den Prozess von der Datenakquise bis hin zur Bereitstellung des fusionierten Ergebnisses. Das Ergebnis besteht aus Zeitschnitten der Umweltbeobachtungen innerhalb der angefragten Grenzen, aus denen eine raum-zeitliche Verteilung beobachteter Arten abgeleitet werden kann.
Entsprechend den Anforderungen an eine GDI, können bei der technischen Umsetzung des Beispiels eine Reihe von OGC Standards eingesetzt werden. Für die Datenbereitstellung sind dies der Sensor Observation Service (SOS, OGC 2012) für gesammelte Umweltbeobachtungen sowie der Web Feature Service (WFS, OGC 2010) für die Bereitstellung administrativer Grenzen. Die Funktionalität zur Datenfusion erfolgt über die OGC WPS Schnittstelle. Das Ergebnis des Fusionsprozesses sind Links zwischen den Ausgangsdaten, welche nach Linked

[2] http://www.w3.org/2003/01/geo/wgs84_pos
[3] http://geovocab.org/geometry
[4] http://www.opengis.net/ont/geosparql

Data Prinzipien in einem RDF (Resource Description Framework) Repository abgelegt werden. Dieses entspricht den Erfordernissen des Semantic Web und bildet gleichzeitig die Grundlage für die Bearbeitung potentielle Abfragen, in deren Folge die gespeicherten Links und darin enthaltenen Informationen zu den Ausgangsdaten genutzt werden können, um entsprechend nutzerspezifische Informationen zu generieren.
Der Prozess der Datenfusion besteht aus mehreren Komponenten, die sich an den in Kapitel 3.1 vorgestellten Teilprozessen orientieren. Zur Ähnlichkeitsmessung zwischen Geoobjekten werden primär die Hausdorff-Distanz (Ähnlichkeit von Geometrien) und die Damerau-Levenshtein-Distanz (Ähnlichkeit von Attributwerten) eingesetzt.
Gegenüber der Umsetzung des abgebildeten Prozesses in klassischen Geoinformationssystemen, bietet der vorgestellte Ansatz eine Reihe von Vorteilen:

- Datenhaltung und -prozessierung laufen dezentral und können nach dem Subsidiaritäts-prinzip parallel entwickelt und gewartet werden.
- Funktionalität der Datenfusion ist durch die Nutzung des OGC WPS Standards nicht an bestehende Systeme gebunden und kann somit frei verwendet und angepasst werden.
- Speicherung der Fusionsergebnisse als Verlinkungen behält die Ursprungsdaten in der originären Form bei und greift erst bei entsprechenden Anfragen auf diese zurück.
- Nutzung von Linked Data ermöglicht die Einbindung weiterer Ressourcen im Web, beispielsweise die Beschreibungen beobachteter Arten aus entsprechenden Fachnetz-werken.
- Zugriff auf die Fusionsergebnisse kann existierende Linked Data Tools nutzen und sowohl über Traversierung von Links als auch die Abfragesprache SPARQL erfolgen.

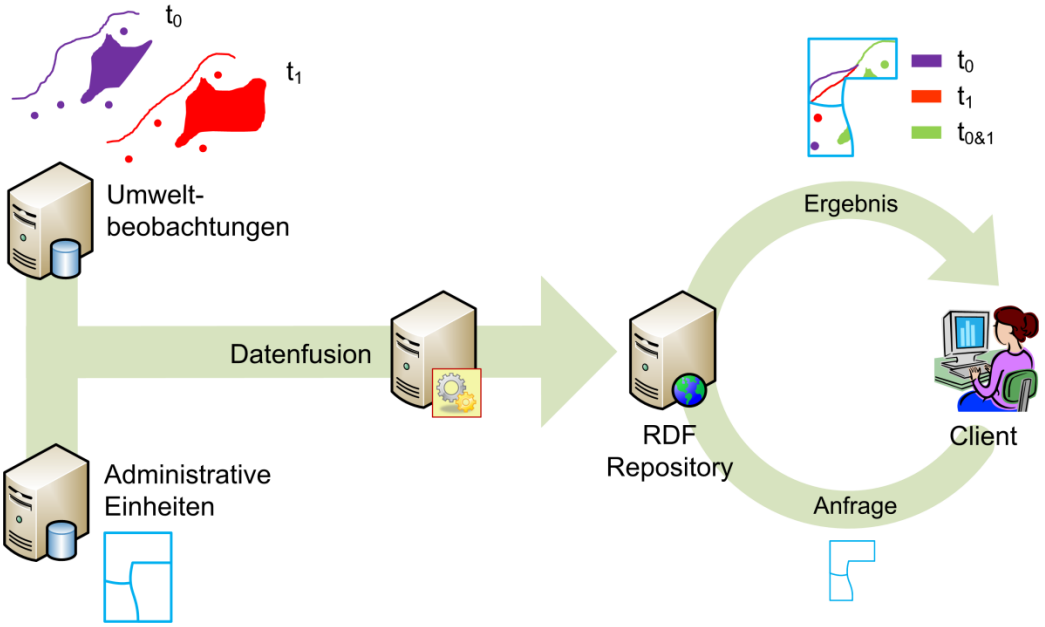

Abb. 3: Prinzipieller Ablauf der Fusion von Umweltdaten mit administrativen Grenzen.

Die beschriebene Umsetzung in COBWEB stellt einen relativ pragmatischen Ansatz dar und ist in der derzeitigen Form lediglich dazu gedacht, die Machbarkeit des Ansatzes zu demonstrieren.

Um das volle Potential der Kombination von GDI und Semantic Web auszunutzen, wird eine Weiterentwicklung in den folgenden Bereichen angestrebt:

- Vollständige Formalisierung und Implementierung möglicher Ergebnisrelationen einer Datenfusion,
- Formalisierung und Implementierung von Konfidenzmaßen zur qualitativen Beschreibung von Ergebnissen aus dem Matchingprozess,
- Erweiterung der Funktionalitäten zur Geodatenfusion in allen, der in Kapitel 3.1 beschriebenen, Teilprozesse,
- Identifizierung und Zusammenführung erfasster Geodaten mit anderen, potentiell ergänzenden Datensätzen im Web,
- Flexible, anwendungsorientierte und zum großen Teil automatisierte Orchestrierung von Prozessierungsdiensten zur Geodatenfusion.

Die Weiterentwicklung im Rahmen des COBWEB Projektes wird zudem Aspekte der Qualitätssicherung von Umweltbeobachtungen aus Crowdsourcing sowie Möglichkeiten zur Anbindung an bestehende GDI, beispielsweise im Kontext von INSPIRE (Infrastructure for Spatial Information in the European Community), mit einbeziehen.

4 Weiterer Forschungsbedarf

Die Entwicklung einer vollautomatisierten, zuverlässigen und anwendungsorientierten Fusion von Geodaten ist ein wichtiger Schritte in Richtung des propagierten Geospatial Semantic Web (EGENHOFER 2002). Die vorgestellte Anwendung demonstriert dabei die Vereinbarkeit von traditionellen GDI und den Prinzipien des Semantic Web. Weiterer Forschungsbedarf in diesem Gebiet besteht insbesondere in den folgenden Gebieten:

- Interoperable Repräsentation und Bereitstellung von Geodaten als Linked Data – beschäftigt sich primär mit der Kodierung sowie Zugriffs- und Speichermöglichkeiten von Geodaten im Semantic Web.
- Kommunikation von GDI Komponenten innerhalb des Semantic Web – zielt auf die Anbindung existierender Dienste zur Bereitstellung, Katalogisierung, Prozessierung und Visualisierung von Geodaten an das Semantic Web.
- Möglichkeiten zur unscharfen Verlinkung von Geodaten – erforderlich für Verknüpfungen und Weiterverarbeitung von Repräsentationen, die nicht exakt, sondern nur partiell oder unter einer gewissen Wahrscheinlichkeit das gleiche realweltliche Objekt abbilden.
- Wartung und Qualitätskontrolle für verlinkte Geodaten – befasst sich mit Möglichkeiten zur Synchronisation und dem Konfliktmanagement zwischen existierenden Datenbestände und deren Verlinkungen im Semantic Web.
- Möglichkeiten zur Informationsgenerierung aus verlinkten Geodaten – zielt auf die effektive Verwertung verlinkter Geodatenbestände, um einen anwendungsspezifischen Mehrwert, beispielsweise zur Entscheidungsunterstützung, abzuleiten.

Zu den genannten Gebieten wird auch im Kontext des COBWEB Projektes weiter geforscht. Entsprechende Ergebnisse werden in weiteren Publikationen oder Projektberichten veröffentlich.

5 Literaturverzeichnis

BIZER, C. & HEATH, T. & BERNERS-LEE, T., 2009: Linked data - the story so far. International Journal on Semantic Web and Information Systems (IJSWIS), **5**(3), S. 1-22.

EGENHOFER, M. J., 2002: Toward the Semantic Geospatial Web. Proceedings of the 10th ACM international symposium on Advances in geographic information systems. S. 1-4.

ERL, T., 2008: SOA – Principles of Service Design. Prentice Hall. ISBN-13: 9780132344821.

GOODCHILD, M.F., 2007: Citizens as sensors: the world of volunteered geography. GeoJournal, **69**(4), S. 211-221.

OPEN GEOSPATIAL CONSORTIUM, OGC, 2007: OpenGIS Web Processing Service, Implementation Standard, Version 1.0.0.

OPEN GEOSPATIAL CONSORTIUM, OGC, 2010: OpenGIS Web Feature Service 2.0 Interface Standard, Implementation Standard, Version 2.0.0.

OPEN GEOSPATIAL CONSORTIUM, OGC, 2012: OGC GeoSPARQL - A Geographic Query Language for RDF Data, Implementation Standard, Version 1.0.

OPEN GEOSPATIAL CONSORTIUM, OGC, 2012: OGC Sensor Observation Service Interface Standard, Implementation Standard, Version 2.0.